

--The present patent application is a Divisional of Application No. 09/289,772, entitled "Extracting Information from Symbolically Compressed Document Images," filed April 8, 1999 and assigned to the corporate assignee of the present invention.--

IN THE CLAIMS

Please cancel claims 2-16, 18-35 and 37-40 without prejudice.

This listing of claims will replace all prior versions, and listings, of claims in the application:

1. (original) A method comprising:

representing an input document image with a sequence of template identifiers to reduce storage consumed by the input document image; and  
replacing the template identifiers with alphabet characters according to language statistics to generate a text string representative of text in the input document image.

Claims 2. – 16. (canceled)

17. A document processing system comprising:

a deciphering module to generate a first text string on a sequence of template identifiers in a first symbolically compressed document image and to generate a second text string based on a sequence of

template identifiers in a second symbolically compressed document

image;

a conditional n-gram module coupled to receive the first and second text

strings from the deciphering module, the conditional n-gram

module being configured to extract n-gram indexing terms from

the first and second text strings based on a predicate condition; and

a comparison module to generate a measure of similarity between the first

and the second symbolically compressed document image based on

the indexing terms extracted by the conditional n-gram module.

Claims 18. – 35. (canceled)

36. An article of manufacture including one or more computer-readable media that embody a program of instructions to generate a text string from an input document image represented by a sequence of template identifiers for the purpose of reducing storage consumed by the input document image, wherein the program of instructions, when executed by one or more processors in the processing system, causes the one or more processors to replace the template identifiers with alphabet characters according to language statistics to generate a text string representative of text in the input document image.

Claims 37. - 40. (canceled)